Underreporting traffic accidents in Malaysia – a sentiment analysis

Zamira Zamzuri[1*]

[1]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Abstract. The underreporting scenario is claimed to be the source of the extra zeros in traffic accident data. This leads to a latter problem in which the fitted statistical model may not be able to produce correct and reliable estimates. Understanding the root of problem as to what is the main cause of the underreporting scenario is essential to assist on the decision making process in traffic accident analysis. In this study, 200 Malaysian drivers were interviewed on their sentiments towards this issue. Their opinions on the causes of underreporting scenario are investigated then assessed using text analyses. First, the Latent Dirichlet Allocation text modelling is employed to find the underlying themes in the reasons of not reporting a traffic accident. Then, the polarity of the topics is measured using a lexicon based sentiment analysis. Results showed that majority Malaysian drivers (80.5%) consider that reporting a minor or non-fatality accident is not important and can be neglected. The decision is due to the fact that of complicated and time consuming reporting process. The drivers are also asked on their opinion after the consequences of underreporting are informed to them. The polarity of their answers shifted to more positive in which 71% drivers will report an accident that occur in the future.

1 Introduction

Traffic accident analysis is vital to society and country as its impact can be huge not only physically but also economically. Based on the report by World Health Organization (WHO), there were 1.35 million road traffic deaths globally in 2016. Currently, road traffic injuries are estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030. Hence it is essential to understand influencing factors and predicting future outcomes in terms of traffic accident frequency and severity. To achieve these objective, statistical analyses have been conducted by utilising the reported traffic accident data as can be found in [1-3]. As mentioned by [4-5], the accident count data often exhibit extra zeros which requires the need of zero-augmented models. The presence of extra zeros in accident count data is associated with the underreporting scenario [6-7]. The underreporting scenario is a situation in which an accident did happen but has not been reported. Other than dealing with the extra zeros through the zero-augmented models, there is also a need to enhance the integration of data in order to

*Corresponding author: zamira@ukm.edu.my

ensure that most of the accidents are reported. In this study, we aim to explore the drivers' opinion on the underreporting issue and what leads them to not reporting an accident.

The issue of underreporting of traffic accidents have been explored before by [8-11]. The underreporting scenario is observed in France with the percentage of unreported accidents is 62%. Other studies on the same issue reveal that the percentage of unreported accidents is relatively high such as 46% in New Zealand [10], 43.8% in Australia [11], 42.5 % in Hong Kong [12] and 20% in the United Kingdom [13]. Levels of severity is identified as the main factor influencing the accident rate as shown in [14] in a meta-analysis of 49 studies conducted in 13 countries. The underreporting rate is 90% for very slight injuries, 75% for slight injuries, 30% for serious injuries and 5% for fatal injuries.

In Malaysia, the issue of unreported traffic accidents has been studied by [15]. In this study, using the police and hospital record in state of Melaka, the police reporting rate is found to be 4.7%. Furthermore, the matching rate between police and hospital records is proportional to the level of injury severity. The database system for the road traffic accidents in the state of Johor Bahru has been developed by [16] since there is a discrepancy in the police report. [17] estimate the proportion of unreported accidents in Malaysia at 58% for slight injuries and 54% for serious injuries. Since there is still a limited number of references on this issue for Malaysia data, this paper intends to estimate the proportion of unreported accidents in Malaysia. By finding the fitted distributions to the proportion of unreported accidents, perhaps a richer information can be obtained in order to shed some light to the actual number of accidents occurred. Furthermore, the distribution fitted to the proportion of unreported accidents data will help in developing a traffic accident model that include underreporting information perhaps to improve the estimation of the actual number of accidents. Another study by [18] reveals that Malays aged between 60 and 69, living in rural area, who are single and are motorcyclists tend not to report the accidents.

In all of the papers mentioned above, the focus is given to the figures of the underreporting rate to show the seriousness of this issue. Furthermore, the accuracy and reliability of the statistical analyses conducted will be improved when a more complete set of data used. However, the reasons on why drivers are not reporting the accidents are only speculated and not specifically quantified through proper investigation and analysis. Hence, this paper aims to fill the gap by mining the opinion of Malaysian drivers towards the underreporting issue. With the advent of technology, unstructured data such as text can be analysed quantitatively as to be shown in this paper. It is vital to understand and quantitatively analysed the reasons of drivers are not reporting a traffic accident. Since the perception established is there is the need on reporting accidents only for major and fatal accidents, hence this perception needs to be changed. As reported by [19], near-misses and minor accidents are crucially important to report in order to avoid major accidents happening in future at the same location with the same environment. Creating awareness on reporting traffic accidents is paramount, hence identifying and further verifying the reasons behind the actions on not reporting accidents are also crucial. Although it seems that the reason of not reporting traffic accidents are known, however, as mentioned earlier; all of these reasons are only discussed and speculated, without any quantification measures and analyses. Hence, it is justified the importance of conducting this study.

2 Data & methodology

In this section, we discuss on the data and methodology for this study.

2.1 Data

Data gathered in this study is based on interview sessions conducted to 200 Malaysian drivers. The drivers are sampled based on the stratification on their demographic profile which are gender, race and age, in order to match with the profile of Malaysian drivers as reported in Transport Statistics Malaysia (2008). Details on the stratified sampling technique can be found in [20]. The main difference of the text data in this study compared to other papers in literature is the text data in this study is extracted from interview sessions, whereas other text data in literature are often scraped from social media platforms or piles of written documents. For text data scraped from the internet, [21] suggests 500 documents are needed in order for the samples to be representative with 95% capture probability. On the other hand, since the data of this study is collected through interview sessions, a time constraint becomes a limitation in order to collect more data. Based on [22], past research often use data gathered from 20 – 30 interviews. Furthermore, the sample selection in this study is based on stratification of the population in which it helps on producing more representative samples [23]. Hence, we conclude that the answer scripts from 200 Malaysian drivers are considered representative samples of the target population.

The interview session is conducted in one session per driver with duration around 15 to 20 minutes. The drivers are recorded on their demographic background and further answer a series of question related to the underreporting issue. The questions projected is shown in Table 1. After the fourth question, the interviewer will explain on the consequences of not reporting an accident even though it is minor. Then, the drivers are asked on their opinion with now knowing on the fact of unreported accidents' consequences.

The interview sessions are recorded audibly and then transformed to a series of transcripts. These transcripts are then analysed using text mining techniques. In this study, we will focus on two questions only before and after the explanation on the consequences. The answers that will be used as data in this study are based on these questions:

Before explanation:

Q1 If you are involved in a minor accident, will you report the accident?

Q2 Why? (if the driver answered 'No' in Q1)

After explanation:

Q6 If you are involved in an accident in the future, will you report the accident? Why?

2.2 Text analysis

Since we are focusing on the text data which is categorized as unstructured, we discuss in this section several steps in text analysis including topic modelling and sentiment analysis

2.2.1 Text data preparation

It is a known fact that text data is unstructured, which make it is messy and tedious to manage compared to the structured data. The text data needs to be cleanse and prepare before the analysis can be conducted. Based on [24], the text data collection known as corpus. Then, the data is being pre-processed by transforming all words to lower case, removing punctuation, symbols, stop-words and numbers. Any additional white space between words are also stripped in this process. Stemming and tokenization are essential steps in preparing the data.

Through stemming, the words are reduced to its root; for example, words such as 'teaching', 'teacher', and 'teaches' are all reduced to 'teach'. Whereas, tokenization refers to splitting the text into a list of token, either as one word or phrases of words.

2.2.2 Topic modelling

The text modelling analysis is employed when we want to find the underlying themes of our text. The Latent Dirichlet Allocation (LDA) model is commonly used in topic modelling. As explained in [25], this model is probabilistic based in which each document is represented as a random mixture of latent topics and each topic is represented as a distribution over fixed set of words. In LDA, each document is considered to be constructed from a number of topics; and each topic has a mixture of words. Hence, the generative process of document based on LDA is considered hierarchical with these layers:

1. For each document *w*, choose *N* from Poisson (μ). *N* is the number of words.
2. Choose θ from Dirichlet distribution with parameter α. Parameter θ here represents the proportion of a topic in the document.
3. For each *N* words in w_n, choose the topic, z_n from a multinomial distribution with parameter θ as generated in 2. Also choose a word, w_n from $P(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic. A corpus is a collection of *M* documents, denoted by *D*.

The probability density function for Poisson, Dirichlet and multinomial distributions are presented in Equations 1 – 3.

$$f(N|\mu) = \frac{e^{-\mu}\mu^{x}}{x!} \tag{1}$$

$$d(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} \tag{2}$$

$$k(z_1, z_2, \dots, z_N|\theta) = \sum_{i=1}^{N}(z_i)!\prod_{i=1}^{N}\frac{\theta_i^{z_i}}{z_i!} \tag{3}$$

Since the LDA is a 3 layers hierarchical Bayesian model, the conditional probability density is proportional to the product of likelihood and priors as shown in Equation 4.

$$P(\theta, z, w|\alpha, \beta) = P(\theta|\alpha)\prod_{n=1}^{N}P(z_n|\theta) + P(w_n|z_n, \beta) \tag{4}$$

The probability of a corpus is determined by taking the product of the marginal probabilities of single documents, as shown in Equation 5.

$$P(D|\alpha, \beta) = \prod_{d=1}^{M}\int P(\theta_d|\alpha)\prod_{n=1}^{N} d\ \sum_{zd_n}P(z_{dn}|\theta_d) + P(w_{dn}|z_{dn}, \beta)d\ \theta_d \tag{5}$$

Number of topics, *k* is often determined first before the LDA model can be fitted to the data steps. There are several approaches suggested in the literature to determine the number of topics such as perplexity and coherence score [26]. [27] mentioned that the decision on the number of topics also depend on humans' judgement to maintain the semantic meaning of the terms. Overall, the mixture of machine automated and human are the best way to determine the number of topics.

Once the topics for each documents are identified along with the related words, the importance of these words to the topic can be computed, β_{ij} represents the probability of the *i*th topic containing the *j*th word. The estimation of parameters in this model is performed using Gibbs sampling in Bayesian framework.

2.2.3 Sentiment analysis

Sentiment analysis is unique in a way it combines the qualitative and quantitative analysis. The context and sentiment in text are extracted and being quantified in this analysis. There are two approaches in sentiment analysis which are lexicon based and machine learning based. In this paper, we only focus on the lexicon based by using the 'bing' lexicon as can be found in the R package, 'tidytext' [28]. This lexicon is chosen due to its simplicity that classifies words into two categories, positive and negative. In general, the sentiment score is computed as the difference between the positive words count and the negative words count.

In this paper, we perform the sentiment analysis to the bigram, which is a pair of words. We are not considering unigram due to the fact that misleading results can be obtained when the phrases contain negation word. In processing the bigram to compute the sentiment score, an additional step is needed in which separating the bigrams that are preceded by negation word such as 'not', 'no', 'never' and 'without'. Then the contribution of these bigrams towards the score is reversed.

3 Findings and discussion

In this section, we present the result of the text analysis conducted to the interview data.

3.1 The drivers and accidents profile

As mentioned in the previous section, the data in this study is based from the interview session conducted on 200 Malaysian drivers. The demographic background is given in Table 1, which is approximately similar representation of Malaysian drivers' population. Based on Table 1, majority of Malaysian drivers are Malays aged 18 -25 who have 6 to 15 years of driving experience and have involved in at least one traffic accident. From these 200 drivers, 859 traffic accidents are recorded. The summary statistics of the accident counts are given in Table 2.

Table 1. The drivers' profile.

Variable		Frequency (%)
Gender	Male	104 (52.0)
	Female	96 (48.0)
Age	18-25	63 (31.5)
	26-35	46 (23.0)
	36-45	37 (18.5)
	46-55	29 (14.5)
	More than 55	25 (12.5)
Race	Malay	139 (69.5)
	Chinese	46 (23.0)
	Indian	14 (7.0)
	Other	1 (0.5)
Driving experience (yrs)	Less than 3	29 (14.5)
	3 to 5	13 (6.5)
	6 to 15	74 (37.0)
	16 to 25	36 (18.0)
	More than 25	48 (24.0)
Accident involvement	Yes	131 (65.5)
	No	69 (34.5)

The accidents' summary statistics in Table 2 shown that only 82.8% of the accidents are reported. Although the proportions of unreported accidents is quite low, around 20%, this figure can be higher if we look at accidents based on the severity levels as reported in [17]. This finding indicates the presence of unreported accidents in Malaysia. Based on Table 2, on average, a Malaysian driver have involved in 4 accidents.

Table 2. The accidents' summary statistics.

Statistics	Total number of accidents	Number of reported accidents
Minimum	0	0
Maximum	10	9
Mean	4.27	3.54
Median	4	3
Standard deviation	1.47	0.94
Total	859	712

3.2 Reasons on why drivers are not reporting minor accidents

In the interview conducted, the first question asked to the respondent is "If you are involved in a minor accident, will you report the accident?". If the respondent answered 'No', then the interviewer asks for the reasons as the second question. The answer scripts of Question 2 are then analysed using topic modelling to find the underlying themes on their reasons. Then, a sentiment analysis is conducted to measure the polarity in the answers. Table 3 depicts the frequency of answers for Question 1 in which obviously we can see that 161 out of 200 drivers won't report a minor accident if the accident happened. This is an alarming figure as 80.5% is a high value in which any accident needs to be reported regardless of its severity.

Table 3. The frequency table for answers in Q1.

Answer	Frequency (%)
Yes	39 (19.5)
No	161 (80.5)

Next, we process the text answers from Question 2 as explained in the methodology section. Out of 161 drivers who answered 'No' for reporting minor accidents, only 148 drivers state the reasons. For these answers, rather than tokenizing the text into a series of single words, we constructed bigram, which is a collection of phrases consists of two words. The reason of conducting text analysis on the bigram compared to unigram is in order to obtain more information and meaningful interpretation from the text extracted. Furthermore, in sentiment analysis, when the unigram is considered, there are words that separated from their negation, hence producing misleading results.

Fig. 1 displays the plot of beta to the top bigrams for each topic. Through the approaches explained before, the number of optimum topics is determined at three. For topic 1, phrases

'insurance claim', 'self-manage' and 'solve problem' are listed which indicates that the drivers are not reporting the minor accidents since there is no need lodge insurance claims and easier to self-managing the accidents. For the second topic, three words are interchangeably listed which are 'report', 'process' and 'complicated' which signifies that the complicated reporting process for traffic accidents demotivates the drivers, hence resulting on them not reporting the accident. The last topic consists of phrases 'waste time', 'serious damage' and 'complicated situation' which shown the sentiment of the drivers towards reporting minor accident is quite negative with the thought that the action is just a waste of time. The drivers think that only accidents with serious damage are need to be reported. Another bigram that caught our attention is 'unnecessarily laziness' that seems to be the negative attitude and personal opinions of the drivers on the issue discussed.

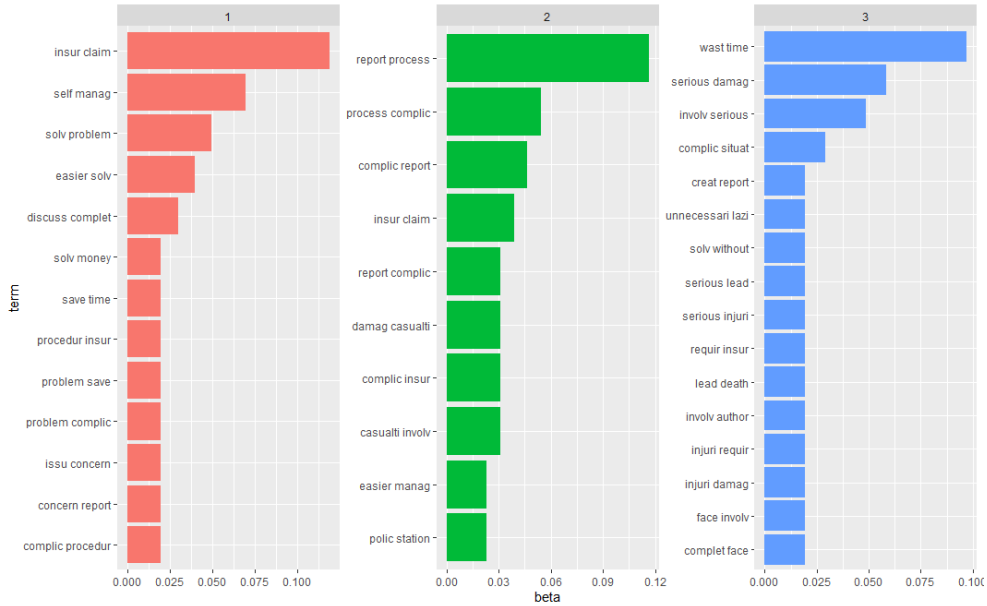

Fig. 1. The important bigrams for the three topics in Q2.

We then proceed by computing the sentiment score based on top 100 bigrams in each topic. Table 4 summarizes the sentiment score in which all topics are relatively negative and most answers are categorized in the first topic which is 'Only report for accidents to lodge insurance claims'. The least negative topic is the first one with the second and third topics have approximately the same score.

Table 4. Summary statistics on the sentiment score for each topic in Q2.

	No of documents	**Total score**	**Average score**
Topic 1	59	-24	-0.407
Topic 2	46	-34	-0.739
Topic 3	43	-33	-0.603

3.3 Will the drivers report an accident in future?

In this section, the drivers have been explained with the consequences of not reporting traffic accidents. With now being an informed driver in the underreporting scenario, the drivers are then asked 'If you are involved in an accident in future, will you report the accident'. Table 5 gives a quick look on the answer in which we can see tremendous changes towards positivity compared to Table 3, with now 71% of the drivers will report the accidents.

Table 5. The frequency table for Q6 answers.

Answer	Frequency (%)
Yes	142 (71)
No	42 (21)
Depends on severity	16 (8)

Similar with the previous section, we conduct the topic modelling and sentiment analyses to the answer scripts on reasons why the drivers will report an accident in the future. Fig. 2 exhibits the word cloud from two-topics LDA model fitted to the data set. Judging from the word cloud, we can see that two main themes in the answers are related to 'insurance claims' and 'safety'. The first topic conveys the information that the drivers will report the accident in order to claim for the insurance. Other than that, the drivers figure that the action of reporting accidents is needed in order to avoid similar risk in future. For the second topic, we can observe that the drivers are more aware on the issue and know that the reporting action is essential for safety purposes. It is also worth to point out that 'other' is also one of the words with high frequency signifying that the drivers are care for others' safety not just themselves.

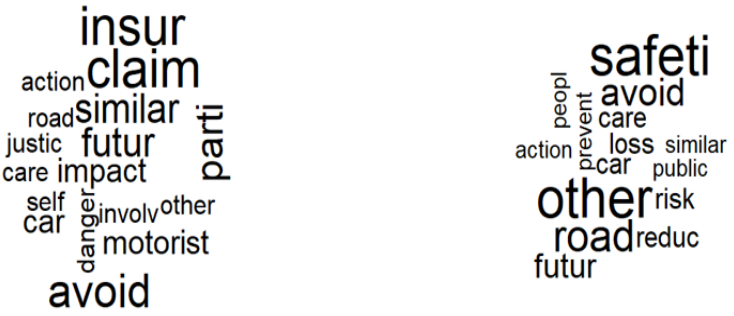

Fig. 2. The word clouds for two topics in Q6.

We then compute the sentiment score for each topic and the results are summarized in Table 6. By examining the scores in detail, we can conclude that Topic 2 is more positive since the theme is about 'safety of others'; however the number of documents categorized in the first topic is higher. Hence, this finding indicates that majority of the drivers will report the accidents in the future for the purpose of insurance claim.

Table 6. Summary statistics on the sentiment scores for each topic in Q6.

	No of documents	Total score	Average score
Topic 1	96	40	0.417
Topic 2	42	22	0.524

4 Conclusions

This paper looks into the causes of underreporting scenario in Malaysia. Malaysian drivers are interviewed to mine their opinion on the issue. Through the session conducted, the drivers are also informed on the negative consequences of not reporting the accidents occurred towards the accuracy and reliability of traffic accident statistical analyses. Response by the drivers are analysed using topic modelling and sentiment analyses. Results reveal that majority drivers won't report a minor accident due to three main factors: cannot file insurance claim, complicated reporting process and waste of time. Out of these three factors, most drivers associated with the 'insurance claim' factor, however 'complicated reporting process' have most negative score in terms of the sentiment. This finding indicates that the traffic accident reporting process needs to be improve in order to encourage the drivers to report an accident.

After the drivers are explained with the consequences of underreporting scenario, they have been asked whether they will report an accident in the future. From this results, the polarity of the drivers' opinion can be seen shifted towards positivity. Two main reasons for this change are 'to lodge insurance claim' and 'safety'. Although more drivers' response are classified into the 'insurance claim' topic compared to 'safety'; the sentiment scores for the second topic is higher. Perhaps, with now being an informed driver, the drivers that participated in this interview session will take action to become more responsible and reporting an accident regardless of the severity level.

There are more to be done in this area. Further analysis can be performed using other lexicons or machine learning based sentiment analysis. Then, the accuracy of these approaches can be compared. The data mining process also can be extended by collecting information from more drivers through social media or any other online platforms. Educating drivers on their role and regulation on the issue of traffic accident especially the underreporting should be a continuous effort.

The author would like to acknowledge Universiti Kebangsaan Malaysia for sponsoring this project under the Research University Grant Scheme (GUP-2018-011).

References

1. K.G. Le, P. Liu, L. Lin, Geo. Spatial Inform. Sci. **23** (2020)
2. M.A. Dereli, S. Erdogan, Transport. Res. A-Pol. **103**, 106 – 117 (2017)
3. M. Schlogl, R. Stutz, G. Laaha, M. Melcher, Accid. Anal. Prev. **127**, 134 – 149 (2019)
4. M.T. Lukusa, F.K.H. Phoa, Stat. Probab. Lett. **158** (2020)
5. P.C. Anastasopoulus, Anal. Methods Accid. Res. **11**, 17 – 32 (2016)
6. Z.H. Zamzuri, M.S. Sapuan, K. Ibrahim, Sains Malays. **47** (2018)

7. D. Lord, S. Washington, J.N. Ivan, Accid. Anal. Prev. **39**, 53 – 57 (2007)
8. J. Alsop, J. Langley, Accid. Anal. Prev. **33(3)** (2001)
9. E. Amoros, J. Martin, L. Laumon, Accid. Anal. Prev. **38** (2006)
10. J. Langley, N. Dow, S. Stephenson, K. Kypri, Inj. Prev. **9**, 376 – 379 (2003)
11. S. Boufous, C. Finch, A. Hayen, A. Williamson, IRMRC (2008)
12. B.P.Y. Loo, K.L.Tsui, Inj. Prev. **13**, 186 – 189 (2007)
13. H. Ward, R. Lyons, R. Thoreau, *Under-reporting of Road Casualties – Phase 1*, in Road Safety Research Report No. 69, London (2006)
14. R. Elvik, A. Mysen, TRR: J. Transport. Res. Board **1665**, 133 – 140 (1999)
15. N. Kamaluddin. M. Abdul Rahman, A. Varhelyi, Int. J Inj. Contr. Saf. Promot. **1**, 52 – 59 (2018)
16. A. Mustaffa, K. Hokao, J. Soc. Transport. Traf. Stud. **3**, 1 – 8 (2012)
17. N.S. Nik Zamri, Z.H. Zamzuri, ASM Sc. J. Special Issue **1**, 1 – 7 (2018)
18. N.S. Nik Zamri, Z.H. Zamzuri , K. Ibrahim, Int. J. Eng. Technol. **7** (2018)
19. S. Jones, C. Kirchsteiger, W. Bjerke, J. Loss Prev. Process. Ind. **12**, 59 – 67 (1999)
20. C. Wu, M.E. Thompson, Sampling theory and practice (Springer, 2020)
21. D. Jucktett, J. Biomed. Inform. **45**, 460 – 470 (2012)
22. K. Vasileiou, J. Barnett, S. Thorpe, T. Young, BMC Med. Res. Methodol. **18** (2018)
23. F. Shi, Dis. Dynam. Transport. Sys. **2015** (2015)
24. V.A. Kozhevnikov, E.S. Pankratoova, Inter. Sc. J. Theoret. App. Sc. **84** (2020)
25. A. Onan, S. Korukoglu, H. Bulut, Inter. J. Comp. Ling. App. **7**, 101 – 119 (2016)
26. W. Zhaou, J.J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, *A heuristic approach to determine an appropriate number of topics in topic modelling*, in Proceedings of the 12th Annual MCBIOS Conference (2015)
27. C. Cai, M.K Linnenluecke, M. Marrone, A.K. Singh, Abacus **55** (2019)
28. M. Hu, B. Liu, *Mining and summarizing customer reviewes 2004*, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2004)